Solution 1: Kernelized Muliclass SVM

(a) We consider the following constrained optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{\theta},\theta_0,\zeta^{(i)}} \quad & \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + C\sum_{i=1}^n \zeta^{(i)} \\
\text{s.t.} \quad & y^{(i)}\left(\left\langle \boldsymbol{\theta}, \phi\left(\mathbf{x}^{(i)}\right)\right\rangle + \theta_0\right) \geq 1 - \zeta^{(i)} \quad \forall\, i \in \{1,\dots,n\}, \\
\text{and} \quad & \zeta^{(i)} \geq 0 \quad \forall\, i \in \{1,\dots,n\}.
\end{aligned}$$

In the optimum, the inequalities will hold with equality (as we minimize the slacks), so $\zeta^{(i)} = 1 - y^{(i)}\left(\left\langle \boldsymbol{\theta}, \phi\left(\mathbf{x}^{(i)}\right)\right\rangle + \theta_0\right)$, but the lowest value $\zeta^{(i)}$ can take is 0 (we do no get a bonus for points beyond the margin on the correct side). So we can rewrite the above:

$$\frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^n \max(1 - y^{(i)}(\boldsymbol{\theta}^\top\phi(\mathbf{x}^{(i)}) - \theta_0), 0).$$

Note that this is essentially the same argument we used in the linear SVM case to write it as the regularized ERM problem with the hinge loss without using a feature map.

(b) Let $\psi(\mathbf{x}, y) = \frac{1}{2}y\phi(x)$, where ϕ is the feature map of the regularized binary ERM problem in (a). Now, if $y \neq y^{(i)}$ it holds that $y = -y^{(i)}$, so that

$$\begin{aligned}
1 &+ \boldsymbol{\theta}^\top\psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^\top\psi(\mathbf{x}^{(i)}, y^{(i)}) \\
&= 1 + \frac{1}{2}y\boldsymbol{\theta}^\top\phi(\mathbf{x}^{(i)}) - \frac{1}{2}y^{(i)}\boldsymbol{\theta}^\top\phi(\mathbf{x}^{(i)}) && \text{(Definition of } \psi) \\
&= 1 + \frac{1}{2}\left(y - y^{(i)}\right)\boldsymbol{\theta}^\top\phi(\mathbf{x}^{(i)}) && \text{(Distributivity)} \\
&= \begin{cases} 1 + \boldsymbol{\theta}^\top\phi(\mathbf{x}^{(i)}), & \text{if } y^{(i)} = -1 \\ 1 - \boldsymbol{\theta}^\top\phi(\mathbf{x}^{(i)}), & \text{if } y^{(i)} = +1 \end{cases} && (\text{Since } y = -y^{(i)}) \\
&= 1 - y^{(i)}\boldsymbol{\theta}^\top\phi(\mathbf{x}^{(i)}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + \boldsymbol{\theta}^\top\psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^\top\psi(\mathbf{x}^{(i)}, y^{(i)}), 0) \\
&= \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^n \max(1 + \boldsymbol{\theta}^\top\psi(\mathbf{x}^{(i)}, -y^{(i)}) - \boldsymbol{\theta}^\top\psi(\mathbf{x}^{(i)}, y^{(i)}), 0) && (y \neq y^{(i)} \text{ implies } y = -y^{(i)}) \\
&= \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^n \max(1 - y^{(i)}\boldsymbol{\theta}^\top\phi(\mathbf{x}^{(i)}), 0).
\end{aligned}$$

(c) The representer theorem tells us that for the solution $\boldsymbol{\theta}^*$ (if it exists) of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ it holds that $\boldsymbol{\theta}^* \in \text{span}\{(\psi(\mathbf{x}^{(i)}, y))_{i=1,\dots,n,y=1,\dots,g}\}$. This means that $\boldsymbol{\theta}$ has to be a linear combination of

$(\psi(\mathbf{x}^{(i)}, y))_{i=1,\dots,n, y=1,\dots,g}$, so that we can write $\boldsymbol{\theta} = \mathbf{X}^\top \boldsymbol{\beta}$ for $\boldsymbol{\beta} \in \mathbb{R}^{ng}$ and

$$\mathbf{X} = \begin{pmatrix} \psi(\mathbf{x}^{(1)}, 1)^\top \\ \psi(\mathbf{x}^{(1)}, 2)^\top \\ \vdots \\ \psi(\mathbf{x}^{(1)}, g)^\top \\ \psi(\mathbf{x}^{(2)}, 1)^\top \\ \vdots \\ \psi(\mathbf{x}^{(n)}, g)^\top \end{pmatrix}.$$

For $\boldsymbol{K} = \mathbf{X}\mathbf{X}^\top$ we obtain that

$$\|\boldsymbol{\theta}\|^2 = \boldsymbol{\theta}^\top \boldsymbol{\theta} = (\mathbf{X}^\top \boldsymbol{\beta})^\top \mathbf{X}^\top \boldsymbol{\beta} = \boldsymbol{\beta}^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{\beta} = \boldsymbol{\beta}^\top \boldsymbol{K} \boldsymbol{\beta}.$$

Further, it holds that

$$\begin{aligned}\boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}) &= \boldsymbol{\beta}^\top \mathbf{X} \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\beta}^\top \mathbf{X} \psi(\mathbf{x}^{(i)}, y^{(i)}) \\ &\overset{(*)}{=} (\boldsymbol{K}\boldsymbol{\beta})_{(i-1)g+y} - (\boldsymbol{K}\boldsymbol{\beta})_{(i-1)g+y^{(i)}}.\end{aligned}$$

In order to see $(*)$ note that $\psi(\mathbf{x}^{(i)}, y)$ corresponds to the $\big((i-1)g+y\big)$-th row of $\mathbf{X}$ and $\psi(\mathbf{x}^{(i)}, y^{(i)})$ corresponds to the $\big((i-1)g+y^{(i)}\big)$-th row of $\mathbf{X}$. Thus, the matrix-vector product $\mathbf{X}\psi(\mathbf{x}^{(i)}, y)$ corresponds to the $\big((i-1)g+y\big)$-th column/row of $\boldsymbol{K} = \mathbf{X}\mathbf{X}^\top$ and the matrix-vector product $\mathbf{X}\psi(\mathbf{x}^{(i)}, y^{(i)})$ corresponds to the $\big((i-1)g+y^{(i)}\big)$-th column/row of $\boldsymbol{K} = \mathbf{X}\mathbf{X}^\top$ (keep in mind that $\boldsymbol{K}$ is symmetric). Finally, computing the inner product of $\boldsymbol{\beta}$ with $\mathbf{X}\psi(\mathbf{x}^{(i)}, y)$ (or $\mathbf{X}\psi(\mathbf{x}^{(i)}, y^{(i)})$) is the same as computing first the matrix-vector product $\boldsymbol{K}\boldsymbol{\beta}$ and then projecting onto the $\big((i-1)g+y\big)$-th entry (or the $\big((i-1)g+y^{(i)}\big)$-th entry).

With this,

$$\begin{aligned}\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0) \\ &= \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{K} \boldsymbol{\beta} + \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max\left(1 + (\boldsymbol{K}\boldsymbol{\beta})_{(i-1)g+y} - (\boldsymbol{K}\boldsymbol{\beta})_{(i-1)g+y^{(i)}}\ ,\ 0\right).\end{aligned}$$

Solution 2: Kernel Trick

The polynomial kernel is defined as

$$k(x, \tilde{x}) = (x^T \tilde{x} + b)^d.$$

Furthermore, assume $x \in \mathbb{R}^2$ and $d = 2$.

(a) Derive the explicit feature map ϕ taking into account that the following equation holds:

$$k(x, \tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle$$

Solution:

$$
\begin{aligned}
k(x,\tilde{x}) &= \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} + b \right)^2 \\
&= (x_1\tilde{x}_1 + x_2\tilde{x}_2 + b)^2 \\
&= (x_1\tilde{x}_1 + x_2\tilde{x}_2)^2 + 2(x_1\tilde{x}_1 + x_2\tilde{x}_2)b + b^2 \\
&= x_1^2\tilde{x}_1^2 + 2x_1\tilde{x}_1x_2\tilde{x}_2 + x_2^2\tilde{x}_2^2 + 2bx_1\tilde{x}_1 + 2bx_2\tilde{x}_2 + b^2 \\
&= \left\langle \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \\ \sqrt{2b}x_1 \\ \sqrt{2b}x_2 \\ b \end{pmatrix}, \begin{pmatrix} \tilde{x}_1^2 \\ \sqrt{2}\tilde{x}_1\tilde{x}_2 \\ \tilde{x}_2^2 \\ \sqrt{2b}\tilde{x}_1 \\ \sqrt{2b}\tilde{x}_2 \\ b \end{pmatrix} \right\rangle \\
&= \langle \phi(x), \phi(\tilde{x}) \rangle
\end{aligned}
$$

(b) Describe the main differences between the kernel method and the explicit feature map.

Solution:

Using the kernel method reduces the computational costs of computing the scalar product in the higher-dimensional features space after calculating the feature map.